Sisteme informationale economice (8)

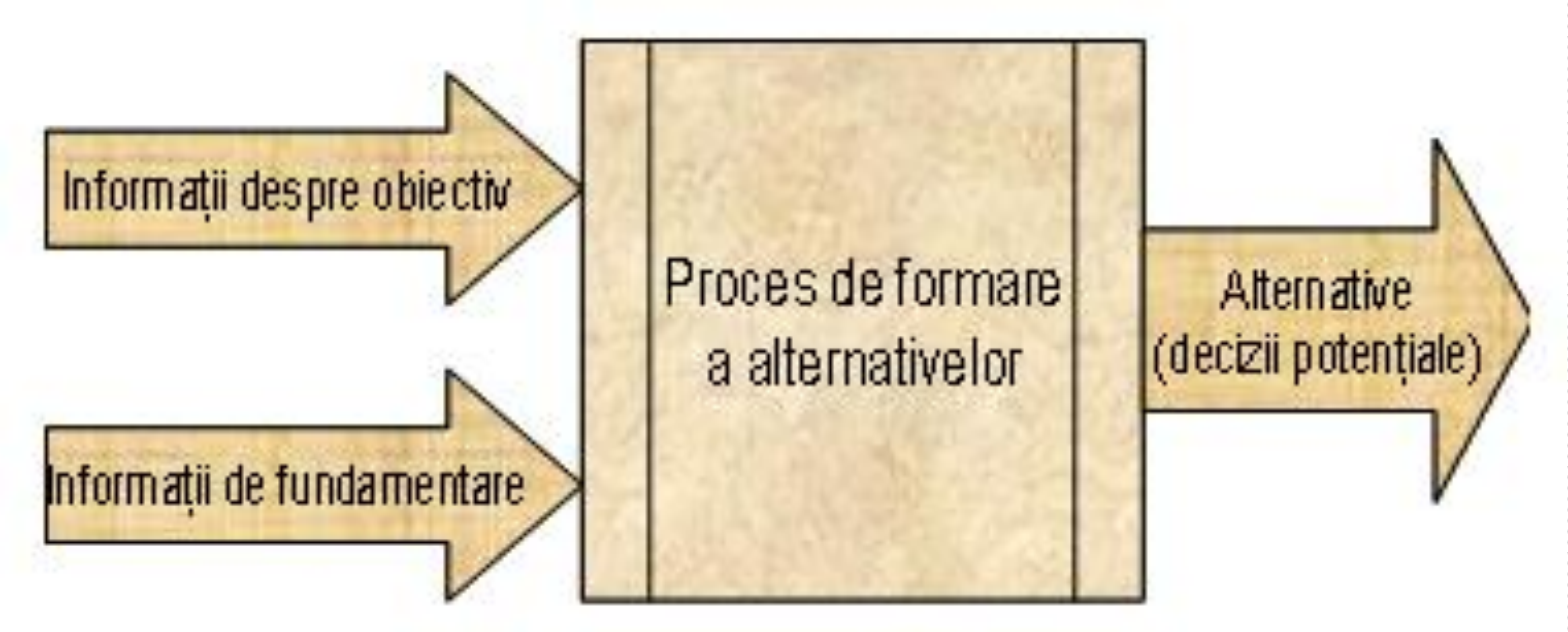
Metode de utilizare a informatiilor in fundamentarea deciziilor

ASE, CSIE, CPE

Structura

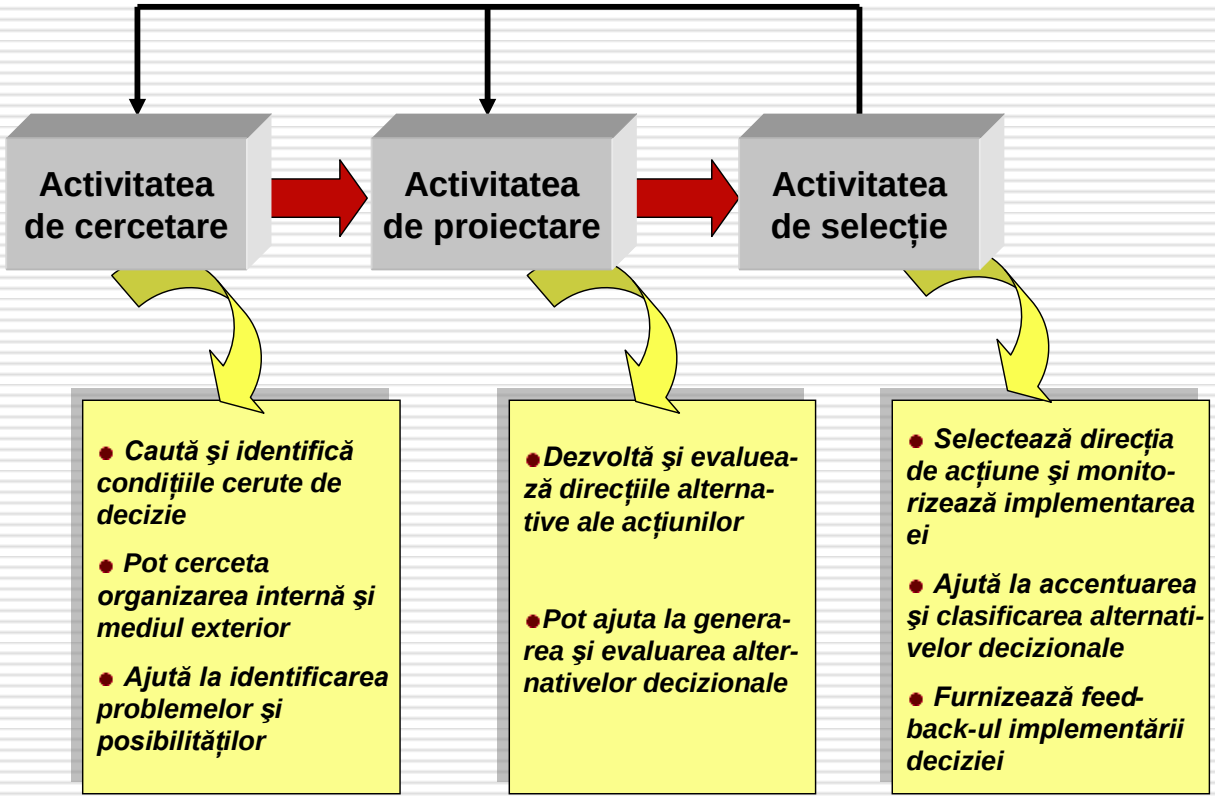
- Modelul procesului de fundamentare a deciziilor
- Modele deductive de fundamentare a deciziilor
- Modele inductive de fundamentare a deciziilor.
- Clasificatorul bayesian naiv

Modelul procesului de fundamentare a deciziilor

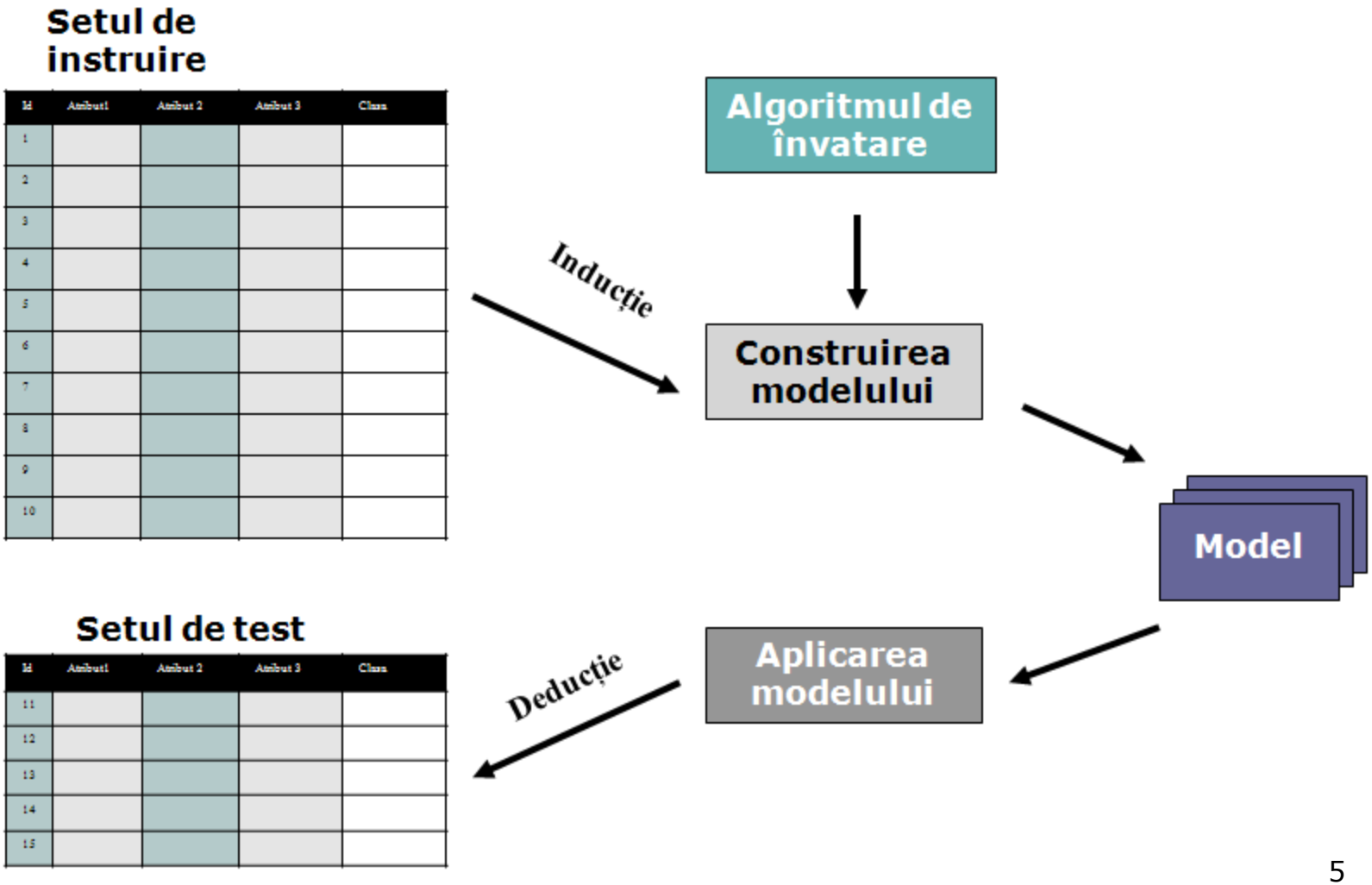


Modele deductive de fundamentare a deciziilor

Modelul Simon, Nobel pentru economie 1978



Modele inductive de fundamentare a deciziilor



Clasificatorul bayesian naiv. Baze teoretice si aplicabilitate

- Teoria rationamentului probabilist definita de Thomas Bayes (regula lui Bayes).
- Regula lui Bayes permite actualizarea probabilitatii ipotezelor pe baza faptelor (evidentei), dupa care sunt alese ipotezele cele mai probabile.
- Metodele de invatare bayesiana sunt utile pentru domeniile cu multe caracteristici.

Exemplu

f1	f2	f3	f4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

Se calculeaza $R_k(n,m)$, ponderea instantelor din clasa m pentru care caracteristica k are valoarea n.

$R_1(1,1) = 1/5$; $R_1(0,1) = 4/5$
$R_1(1,0) = 5/5$; $R_1(0,0) = 0/5$

$R_2(1,1) = 1/5$; $R_2(0,1) = 4/5$
$R_2(1,0) = 2/5$; $R_2(0,0) = 3/5$

$R_3(1,1) = 4/5$; $R_3(0,1) = 1/5$
$R_3(1,0) = 1/5$; $R_3(0,0) = 4/5$

$R_4(1,1) = 2/5$; $R_4(0,1) = 3/5$
$R_4(1,0) = 4/5$; $R_4(0,0) = 1/5$

Avand un vector X de caracteristici (vector de input), clasa in care se repartizeaza X (valorea y care se asociaza lui X) se determina cu ajutorul valorilor R.

Presupunem: X=<0,0,1,1>

- Scorul lui X, daca presupunem ca X apartine clasei 1:

  $S(1) = R_1(0,1)*R_2(0,1)*R_3(1,1)*R_4(1,1)$

  $S(1) = 0.205$

- Scorul lui X, daca presupunem ca X apartine clasei 0:

  $S(0) = R_1(0,0)*R_2(0,0)*R_3(1,0)*R_4(1,0)$

  $S(0) = 0$

$S(1) > S(0) \Rightarrow$ X se clasifica in clasa 1 (y=1).

Algoritmul de clasificare – faza de instruire

Pe baza instantelor de instruire se calculeaza:

$$\mathbf{R_j}(1,1) = \frac{\mathbf{count}(\mathbf{x_j^i} = 1 \wedge \mathbf{y^i} = 1)}{\mathbf{count}(\mathbf{y^i} = 1)}$$

$$\mathbf{R_j}(0,1) = 1 - \mathbf{R_j}(1,1)$$

$$\mathbf{R_j}(1,0) = \frac{\mathbf{count}(\mathbf{x_j^i} = 1 \wedge \mathbf{y^i} = 0)}{\mathbf{count}(\mathbf{y^i} = 0)}$$

$$\mathbf{R_j}(0,0) = 1 - \mathbf{R_j}(1,0)$$

Algoritmul de clasificare – faza de predictie

Fiind dat un vector X, se calculeaza scorul:

$$S(1) = \prod_j \begin{cases} R_j(1,1) \; daca \;\; x_j = 1 \\ R_j(0,1) \quad altfel \end{cases}$$

$$S(0) = \prod_j \begin{cases} R_j(1,0) \; daca \;\; x_j = 1 \\ R_j(0,0) \quad altfel \end{cases}$$

y=1, daca S(1) > S(0)

Corectia Laplace

- Instantele de instruire reflecta doar partial realitatea.
- Daca in cadrul instantelor nu exista valori de caracteristici prezente la anumite clase, R va primi valoare 0.
- Se recomanda evitarea valorilor 0/1 pentru R.

Modificarea formulelor de calcul pentru R

$$R_j(1,1) = \frac{count(x_j^i = 1 \wedge y^i = 1) + 1}{count(y^i = 1) + 2}$$

$$\mathbf{R_j}(0,1) = 1 - \mathbf{R_j}(1,1)$$

$$R_j(1,0) = \frac{count(x_j^i = 1 \wedge y^i = 0) + 1}{count(y^i = 0) + 2}$$

$$\mathbf{R_j}(0,0) = 1 - \mathbf{R_j}(1,0)$$

Exemplu revizuit

f1	f2	f3	f4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

$R_1(1,1) = 2/7$; $R_1(0,1) = 5/7$
$R_1(1,0) = 6/7$; $R_1(0,0) = 1/7$

$R_2(1,1) = 2/7$; $R_2(0,1) = 5/7$
$R_2(1,0) = 3/7$; $R_2(0,0) = 4/7$

$R_3(1,1) = 5/7$; $R_3(0,1) = 2/7$
$R_3(1,0) = 2/7$; $R_3(0,0) = 5/7$

$R_4(1,1) = 3/7$; $R_4(0,1) = 4/7$
$R_4(1,0) = 5/7$; $R_4(0,0) = 2/7$

$X = <0,0,1,1>$
$S(1) = 0.156$
$S(0) = 0.017$

$S(1) > S(0) \Rightarrow$ X se clasifica in clasa 1(y=1).

Linearizarea algoritmului

- Precizia calculelor poate fi afectata de cresterea numarului de probabilitati care se inmultesc.

- Se linearizeaza prin logaritmare:

$$\log S(1) = \sum_j \begin{cases} \log R_j(1,1), daca\ x_j = 1 \\ \log R_j(0,1), altfel \end{cases}$$

$$\log S(0) = \sum_j \begin{cases} \log R_j(1,0), daca\ x_j = 1 \\ \log R_j(0,0), altfel \end{cases}$$

Y=1, daca log S(1) > log S(0)

Clasificatorul bayesian naiv

y=1 daca:

$$\prod_j \alpha_j x_j + (1-\alpha_j)(1-x_j) > \prod_j \beta_j x_j + (1-\beta_j)(1-x_j)$$

unde: $\alpha_1, \ldots \alpha_n, \beta_1, \ldots, \beta_n$ sunt valorile R.

Avantaje si dezavantaje

- Determinarea parametrilor R nu urmareste minimizarea erorii si utilizeaza o metoda rapida de calcul.

- Fiecare caracteristica joaca un anumit rol in clasificare (voteaza pentru clasa 1 sau 0). Ponderea caracteristicii (ponderea votului) este:

$$\log \frac{\alpha_j}{1-\alpha_j} - \log \frac{\beta_j}{1-\beta_j}$$

- Influenta fiecarei caracteristici asupra rezultatului clasificarii se poate stabili in mod independent, fiind ulterior combinata cu influenta celorlalte caracteristici (prin multiplicare).

Limitari: exemplul XOR

f1	f2	f3	f4	y
0	1	1	0	0
1	0	1	0	0
1	0	0	1	0
0	1	0	1	0
1	1	1	0	1
0	0	0	1	1

$R_1(1,1) = 2/4$; $R_1(0,1) = 2/4$
$R_1(1,0) = 3/6$; $R_1(0,0) = 3/6$

$R_2(1,1) = 2/4$; $R_2(0,1) = 2/4$
$R_2(1,0) = 3/6$; $R_2(0,0) = 3/6$

$R_3(1,1) = 2/4$; $R_3(0,1) = 2/4$
$R_3(1,0) = 3/6$; $R_3(0,0) = 3/6$

$R_4(1,1) = 2/4$; $R_4(0,1) = 2/4$
$R_4(1,0) = 3/6$; $R_4(0,0) = 3/6$

Pentru orice X nou

$S(1) = 0.625$

$S(0) = 0.625$

Inferenta probabilista

- Se considera caracteristicile si rezultatul clasificarii ca reprezentand variabile aleatoare.

- Algoritmul de clasificare – faza de invatare:

  $Pr(Y=1|f_1, \ldots, f_n)$

- Algoritmul de clasificare – faza de predictie:

  Fiind data o instanta noua, se calculeaza Pr pentru output 1. Daca Pr>0.5 se previzioneaza 1, altfel 0.

Estimarea distributiei Pr(Y=1|f_1, … , f_n)

- Regula lui Bayes:

$$\Pr(A \mid B) = \Pr(B \mid A)\frac{\Pr(A)}{\Pr(B)}$$

- Pentru problema de clasificare:

P(Y=1|f_1, … ,fn)=P(f_1, … ,fn|Y=1)P(Y=1)/P(f_1, … ,fn)

Pr(f1, … , fn) este independenta de Y

P(Y=1) = probabilitatea apriori (o putem considera 0.5)

Trebuie sa ne concentram asupra: Pr(f1, … , fn|Y=1)

- Ipoteza algoritmului: Independenta variabilelor

$$\Pr(f_1,\ldots,f_n \mid Y=1) = \prod_j \Pr(f_j \mid Y=1)$$

Algoritmul de clasificare – faza de instruire

$$R(f_1 = 1 \mid Y = 1) = \frac{count(x_j^i = 1 \wedge y^i = 1)}{count(y^i = 1)}$$

$$R(f_j = 0 \mid Y = 1) = 1 - R(f_j = 1 \mid Y = 1)$$

$$R(f_j = 1 \mid Y = 0) = \frac{count(x_j^i = 1 \wedge y^i = 0)}{count(y^i = 0)}$$

$$R(f_j = 0 \mid Y = 0) = 1 - R(f_j = 1 \mid Y = 0)$$

Algoritmul de clasificare – faza de predictie

Fiind dat un vector x, se calculeaza scorul:

$$S(x_1 \ldots x_n \mid Y=1) = \prod_j \begin{cases} R(f_j = 1 \mid Y = 1) \; daca \;\; x_j = 1 \\ R(f_j = 0 \mid Y = 1) \quad altfel \end{cases}$$

$$S(x_1 \ldots x_n \mid Y=0) = \prod_j \begin{cases} R(f_j = 1 \mid Y = 0) \; daca \;\; x_j = 1 \\ R(f_j = 0 \mid Y = 0) \quad altfel \end{cases}$$

y=1, daca $S(x_1 \ldots x_n|Y=1) > S(x_1 \ldots x_n|Y=0)$

Studiu de caz

Clasificatorul Bayesian aplicat in identificarea spam-urilor (filtrarea mesajelor)

Ce este spam-ul?

Modalități de combatere spam.

Avantajele filtrului Bayesian.